Is stance detection topic-independent and cross-topic generalizable? - A reproduction study

Reuver, M.; Verberne, S.; Morante, R.; Fokkens, A.; Al-Khatib, K.; Hou, Y.; Stede, M.

Is Stance Detection Topic-Independent and Cross-topic Generalizable? – A Reproduction Study

Myrthe Reuver[1] **Suzan Verberne**[3] **Roser Morante**[1] **Antske Fokkens**[1,2]
[1]Computation Linguistics and Text Mining Lab, Vrije Universiteit Amsterdam
[2] Dept. of Mathematics and Computer Science, Eindhoven University of Technology
[3] Leiden Institute of Advanced Computer Science, Leiden University
{myrthe.reuver,antske.fokkens,r.morantevallejo}@vu.nl
s.verberne@liacs.leidenuniv.nl

Abstract

Cross-topic stance detection is the task to automatically detect stances (pro, against, or neutral) on unseen topics. We successfully reproduce state-of-the-art cross-topic stance detection work (Reimers et al., 2019), and systematically analyze its reproducibility. Our attention then turns to the cross-topic aspect of this work, and the specificity of topics in terms of vocabulary and socio-cultural context. We ask: To what extent is stance detection topic-independent and generalizable across topics? We compare the model's performance on various unseen topics, and find topic (e.g. abortion, cloning), class (e.g. pro, con), and their interaction affecting the model's performance. We conclude that investigating performance on different topics, and addressing topic-specific vocabulary and context, is a future avenue for cross-topic stance detection.

1 Introduction

(Online) debate has long been studied and modelled by computational linguistics with argument mining tasks such as stance detection. Stance detection is the task of automatically identifying the stance (agreeing, disagreeing, and/or neutral) of a text towards a debated topic or issue (Küçük and Can, 2020; Schiller et al., 2021).[1] Its use-cases increasingly relate to online information environments and societal challenges, such as argument search (Stab et al., 2018), fake news identification (Hanselowski et al., 2018), or diversifying stances in a news recommender (Reuver et al., 2021).

Cross-topic stance detection models should thus be able to deal with the quickly changing landscape of (online) public debate, where new topics and issues appear all the time. As Schlangen (2021) described in his recent paper on natural language processing (NLP) methodology, generalization is a main goal of computational linguistics. A computational model (e.g. a stance detection model) should learn task capabilities beyond one set of datapoints, in our case: beyond one debate topic.

Cross-topic stance detection is especially challenging because generalization to a new discussion topic is not trivial. Expressing stances is inherently socio-cultural behavior (Du Bois, 2007), where social actors place themselves and targets on dimensions in the socio-cultural field. This also comes with very topic-specific word use (Somasundaran and Wiebe, 2009; Wei and Mao, 2019). For instance, an *against* abortion argument might be expressed indirectly with a 'pro-life' expression, and someone aware of the socio-cultural context of this debate will be able to recognize this. Knowledge from other debate topics such as *gun control* may not be useful, since the debate strategies might change per topic. Despite these fundamental challenges, pre-trained Transformer models show promising results on cross-topic argument classification (Reimers et al., 2019; Schiller et al., 2021).

In this paper, we investigate the ability of cross-topic stance detection approaches to generalize to different debate topics. Our question is: *To what extent is stance detection topic-independent and generalizable across topics?*

Our contributions are threefold. We first complete a reproduction of state-of-the-art cross-topic stance detection work (Reimers et al., 2019), as reproduction has repeatedly shown to be relevant for NLP (Fokkens et al., 2013; Cohen et al., 2018; Belz et al., 2021). The reproduction is largely successful: we obtain similar numeric results. Secondly, we investigate the topic-specific performance of this model, and conclude that BERT's performance fluctuates on different topics. Additionally, we find that a bag-of-words-based SVM model can rival its performance for some topics. Thirdly, we relate this to the nature of the stance detection modelling

[1]There is a wide array of datasets, definitions, and operationalizations of stance detection and classification, and recently Schiller et al. (2021) gave a great overview in their Section 2, as do Küçük and Can (2020) in their survey.

Proceedings of The 8th Workshop on Argument Mining, pages 46–56
Punta Cana, Dominican Republic, November 10–11, 2021.

task, which is inherently more connected to socio-cultural aspects and topic-specific differences than related tasks such as sentiment analysis.

This paper is organized as follows. Section 2 discusses earlier work on stance detection, and specifically generalizability across topics. Section 3 presents the reproduction results. Section 4 adds additional, topic-specific analyses of the classification performance and a bag-of-words-based model to find topic-(in)dependent features. This is followed by our conclusions in Section 5.

2 Background

2.1 Definition of Stance Detection

Stance detection is a long-established task in computational linguistics. Küçük and Can (2020) identify its most commonly used task definition: "For an input in the form of a piece of text and a target pair, stance detection is a classification problem where the stance of the author of the text is sought in the form of a category label from this set: Favor, Against, Neither." (Küçük and Can, 2020, p. 2).[2] The number of stance classes can vary from 2 to 4, e.g. by adding 'comment' and 'query' next to 'for' and 'against' (Schiller et al., 2021). Küçük and Can (2020) emphasize that this computational definition is built upon the linguistic phenomenon of actors communicating their evaluation of targets, by which they place themselves and their targets on "dimensions in the sociocultural field" (Du Bois, 2007, p. 163). Current work focuses mostly on debates deemed controversial in the U.S. socio-political domain, such as abortion and gun control.

2.2 Prior work

Early work on stance detection focused on parliamentary debates and longer texts (Thomas et al., 2006). Since Mohammad et al. (2016)'s stance detection shared task, Twitter has attracted a lot of attention in NLP work on stance detection (Zhu et al., 2019; Darwish et al., 2020; Hossain et al., 2020). Others addressed stance detection in the news domain, with (fake) news headlines (Ferreira and Vlachos, 2016; Hanselowski et al., 2018), disinformation (Hardalov et al., 2021) and user comments on news websites (Bošnjak and Karan, 2019).

Feature-based approaches have largely been replaced by end-to-end neural models. Stance detection has seen a performance increase due to pre-trained Transformer models such as BERT (Devlin et al., 2019). Reimers et al. (2019) reported .20 point F1 improvement over an LSTM baseline with a pre-trained BERT model. Combining multiple stance detection datasets in fine-tuning such a pre-trained Transformer again led to a performance increase, though this model lacks robustness against slight test set manipulations (Schiller et al., 2021).

2.3 Generalization to new topics

Recent work has specifically worked on identifying stances on topics not seen in training. Reimers et al. (2019) train their model on detecting stances and arguments for unseen topics. In their approach however, they treat all topics and stances on these topics as similar and comparable, and report one averaged evaluation metric over topics.

Earlier work (Somasundaran and Wiebe, 2009) already established that ideological stances on topics deemed controversial, such as gay rights, are expressed in a topic-specific manner. Topic-specific features were more informative for SVM models than more topic-independent features.

In more recent work, Wei and Mao (2019) instead specifically focus on how generalizable certain topics are for transferring knowledge to new topics on stance detection. Some Twitter discussion topics seem to share a latent, underlying topic (e.g. both feminism and abortion have the latent topic of equality). In a (latent) topic-enhanced multi-layer perceptron (MLP) model with RNN representation of the tweet, the model indeed uses shared vocabulary between the related topics.

Allaway et al. (2021) notice that earlier work, when considering training on some topics and testing on others, incorporates topic-relatedness. Unlike these other studies however, Allaway et al. (2021, p. 4756) "do not assume a relationship between training and test topics" as a fairer test of robustness. Results they present do show that stance detection is related to topic, but their efforts go to finding topic-invariant stance representations, which improves the generalizability of their model. Their consideration of topic similarity shows that topic difference is very relevant to stance detection.

ALDayel and Magdy (2021) describe in their survey how several studies (Klebanov et al., 2010; Zhu et al., 2019; Darwish et al., 2020) show that texts pro or against an issue use different vocabularies (e.g. using 'pro-life' when expressing a stance

[2] We would like to note that the stance expressed in a text unit does not have to be the stance of an author, e.g. in cases where someone is writing a piece in which they express or quote someone else's opinion.

against abortion). Some of these studies attempt to leverage these vocabularies to generalize across similar topics. Recent work has looked into generalizing stance detection across datasets, task definitions, and domains (Schiller et al., 2021), in which topic-specific performance is not mentioned.

A recent approach to topic-specificity in stance detection is task adaptation. Stein et al. (2021) acknowledge that stance detection usually requires knowledge about the topic of discussion, which is not available for unseen topics. They approach this problem by changing the task to "same-side stance classification", in which a model is trained to classify whether two arguments either have the same or a different stance. This reduces the model's leaning on topic-specific pro- and con-vocabulary, while still being able to separate different stances on the same topic. The best approach to this adapted task on a dedicated leaderboard[3] receives an F1 of .72 in the cross-topic setting with a fine-tuned BERT model (Ollinger et al., 2020).

Our current work adds the discussion of topic difference and topic specificity to state-of-the-art stance detection results. That is, earlier bag-of-words-based work considered lexical specificity of different topics for stance detection, and we add that into the discussion for the current state of the art: pre-trained, end-to-end neural models.

3 Reproduction Experiments

Reimers et al. (2019) apply their approach of cross-topic claim classification to two datasets: the *UKP Sentential Argument Mining Corpus* (Stab et al., 2018) ('the UKP dataset') and the *IBM Debater: Evidence Sentences* dataset (Shnarch et al., 2018) ('the IBM dataset'). We focus on the UKP Dataset, since the IBM Debater dataset has no 'pro' and 'con' class, but rather 'evidence' and 'no evidence' (and our focus is on stance detection). As a second step after stance classification, the authors also attempt to cluster similar arguments within the same topic in a cross-topic training setting. We do not replicate this component, but instead dive deeper into the classification results.

We adopt the definition of reproduction by Belz et al. (2021): repeating the experiments as described in the earlier study, with the exact same data and software. We analyze our reproduced results according to the three dimensions of reproduction proposed by Cohen et al. (2018): whether we find either the same or different (1) (numeric) values, (2) findings, and (3) conclusions as the earlier study.[4] Reproducing the same **values** means obtaining the same numeric results from a specific experiment. Experiments involving fine-tuning on BERT are non-deterministic. We therefore consider the metric fully reproduced if the original result lies within two standard deviations (stdevs) from our result, obtained from 10 random seeds.[5] The same **finding** means that the relation between the values associated with two or more dependent variables is the same, i.e. a system that outperformed another in the original study also does this in the reproduced study. The **conclusion** is the same when the broader implication of findings and values is the same. Conclusions are thus a matter of interpretation. As such, the same findings can lead to different conclusions and conclusions are, contrary to findings, not repeatable (Cohen et al., 2018). This section focuses on the repeatable components of reproducing a study: the values and the findings. We address the conclusions using our more detailed analyses in Section 4.

3.1 Dataset Description

The UKP dataset (Stab et al., 2018) consists of 25,492 argument sentences from 400 Internet texts (from essays to news texts) on 8 topics. The dataset designer's definition of claim is "a span of text expressing evidence or reasoning that can be used to either support or oppose a given topic" (Stab et al., 2018, p. 3665). They define topic as "some matter of controversy for which there is an obvious polarity for possible outcomes" (Stab et al., 2018, p. 3665), and map this polarity to a text expressing one of two classes: for or against the use, adoption, or idea of the topic under discussion. A third class is 'no argument' to the topic under discussion, i.e. the text span falls outside of this polarity.

The 8 topics in the dataset were randomly chosen from online lists of controversial topics on discussion websites (Stab et al., 2018, p. 3666). Specifically, these topics are *abortion*, *cloning*, *death penalty*, *gun control*, *marijuana legalization*, *minimum wage*, *nuclear energy* and *school uniforms*. The stance classes (pro, con, and no argument) were annotated by two argument mining experts

[3] https://webis.de/events/sameside-19/, Accessed on the 22th of September 2021.

[4] For reasons of clarity, we present these dimensions in reverse order compared to Cohen et al. (2018).

[5] The paper we reproduce, Reimers et al. (2019), does not provide model performance standard deviation over seeds.

and seven U.S. crowdworkers. The distribution of the dataset for different topics is shown in Table 1.

In Stab et al. (2018) we see a difference in agreement on stance classes in different topics, especially between expert and crowd. The topic achieving the highest agreement between crowd worker and expert is *school uniforms* ($\kappa = .889$), and the lowest is *death penalty* ($\kappa = .576$). The standard deviation over topics is .08 for expert–expert coded data and .16 for expert–crowd coded, both with a mean of $\kappa = .72$.

3.2 Obtaining the Data

The UKP Dataset is not available online due to copyright concerns, but there is a scraping script with archived hyperlinks available on Reimers et al. (2019)'s GitHub page. We ran this script with all specifications given. The scraping script was able to return all claims on 6 of the 8 topics. The topics for which not all claims were detected were *nuclear energy* and *minimum wage*. We then instead obtained the complete datafiles from the authors.[6]

3.3 Training and Evaluation Method

Reimers et al. (2019) use the training method described in Stab et al. (2018). Each topic is split into a training (70%), development (10%), and test split (20%). Training is done on the training splits of 7 topics, tuned on the development split (10%) of these 7 topics, and finally evaluated on the test split (20%) of the held-out 8th topic. They do this for each of the 8 topics (holding out a different topic each time), then apply this procedure for 10 different random seeds on a GPU. Evaluation is assessed with macro F1, averaged over all topics and all random seeds. Their best performing model is a fine-tuned BERT-large model (Devlin et al., 2019), but with only minor improvement over BERT-base.

We use the same training set-up and BERT models for our reproduction. For training, we use the author's code with Python3.8 on a single NVIDIA GeForce RTX 2080 Ti GPU. Our learning rate is 2e-5 for both models, as in Reimers et al. (2019).[7]

We additionally train a non-BERT model (a Support Vector Machine (SVM) with tf-idf features) in the same hold-one-topic-out manner. Tf-idf-based approaches have shown quite solid performance on stance detection in prior work (Riedel et al., 2017). This model is deterministic and is thus not run with multiple seeds. It is run with Python3.9 and the sklearn package. The SVM is intended for the feature analysis in Section 4.3, but we present the performance of this model also in Table 2 and the following section.

3.4 Results of Reproduction

BERT-base Table 2 shows that mean performance over the 3 classes ('pro', 'con', or 'no argument') is F1 = .617 (stdev over 10 seeds = .006). Reimers et al. (2019)'s reported result (F1 = .613) lies within 1 stdev from this result.

BERT-large Mean performance over all topics and stance classes is F1 = .596 (stdev over 10 seeds = .043). The performance reported in Reimers et al. (2019) is F1 = .633, which lies within 2 stdev of our result. However, our stdev is relatively high due to high variance of performance over different seeds, with half of our seeds performing noticeably lower than even BERT-base.[8] For the other 5 seeds, the model performed better (F1 = .636, stdev = .007), and within one (much smaller) stdev of the performance reported in Reimers et al. (2019).

SVM+tf-idf (non-BERT model) This model performs at F1 = .517 averaged over the held-out topics and three classes ('pro', 'con', and 'no argument'), see Table 2. This outperforms by .10 points in F1 the best performing LSTM-based architecture presented in Stab et al. (2018) (F1 = .424), a baseline in Reimers et al. (2019). Their performance improvement of the BERT model over LSTM was .20 in F1. Comparing our SVM model to BERT, we find a smaller improvement over a non-BERT model: .10 F1 improvement for BERT-base (F1 = .617). Our BERT models still outperform our

[6] These files revealed that the scraping script broke down in the *minimum wage* topic due to one specific claim that was archived, but could not be retrieved. "Despite the inevitable negative outcomes that will surely result from a $ 15 minimum wage – we 've already seen negative effects in Seattle 's restaurant industry – politicians and unions seem intent on engaging in an activity that could be described as an "economic death wish." We speculate this claim could possibly not be retrieved due to it containing the dollar sign, https://web.archive.org/web/20160217041546/http://www.aei.org:80/publication/ten-reasons-economists-object-to-the-minimum-wage/

[7] All our code can be found in the following GitHub repository: https://github.com/myrthereuver/claims-reproduction.

[8] Our large variance in performance over seeds is due to each seed fine-tuning the model 8 times (once for each topic). The 5 unevenly performing seeds each under-perform on a different topic (F1 < .50) due to only assigning the majority class ('no argument'). Other topics in these 5 seeds do outperform BERT-base.

		abortion	cloning	death penalty	gun control	marijuana legalization	minimum wage	nuclear energy	school uniform	all
train	*pro*	490	508	316	566	422	414	436	392	3.544
	con	591	604	789	479	450	396	613	525	4.447
	no arg	1.746	1.075	1.522	1.359	908	968	1.524	1.248	10.350
dev	*pro*	54	56	38	63	47	46	48	44	396
	con	66	67	90	53	50	44	68	58	496
	no arg	195	120	165	152	101	108	170	139	1.150
test	*pro*	136	142	103	158	118	116	122	109	1.004
	con	165	168	232	133	126	111	171	146	1.252
	no arg	486	299	396	378	253	270	424	347	2.853

Table 1: Distribution of the UKP data over topics and over training (70%), test (20%), and validation (10%) sets.

Model	**UKP Dataset**				
mean (stdev) 10 seeds	F1	P pro	P con	R pro	R con
Reimers et al. (2019) biclstm+BERT	.424	.267	.389	.281	.403
Reimers et al. (2019) BERT base	.613 (-)	.505 (-)	.531 (-)	.470 (-)	.576 (-)
Reimers et al. (2019) BERT large	**.633** (-)	.554 (-)	.584 (-)	.505 (-)	.560 (-)
SVM+tf-idf	.517	.418	.460	.414	.423
Reproduction BERT-base	**.617 (.006)**	.519 (.011)	.538 (.007)	.464 (.029)	.581 (.019)
Repr. BERT-large - all seeds	.596 (.043)	.483 (.057)	.527 (.057)	.464 (.058)	.516 (.063)
Repr. BERT-large - 5 evenly performing seeds	.636 (.007)	.532 (.014)	.578 (.016)	.515 (.016)	.567 (.022)

Table 2: Reproduction results Reimers et al. (2019). The fourth row shows our non-BERT model (an SVM) beating their LSTM baseline, and the fourth and fifth row show the results of our BERT reproductions. The sixth row shows an average BERT-large performance without the 5 seeds that considerably under-performed for one topic.

non-BERT model, as in Reimers et al. (2019). Our SVM result does fall within 2 stdevs of BERT-large, but this is due to BERT-large's substantial stdev due to a steep drop in performance for half of the seeds.

3.5 Conclusion of reproduction

Reimers et al. (2019)'s results are reproducible in the sense the first dimension of reproducibility (Cohen et al., 2018): the originally reported numeric **values** fell within 2 stdevs of our reproduced results for both BERT-base and BERT-large. For BERT-base and 5 of the 10 seeds in BERT-large, we obtained a precision, recall, and F1 that are very similar to the original study.

The results are also reproducible in four of the five reproducibility aspects identified by Fokkens et al. (2013): under-descriptions of preprocessing, experimental set-up, versioning, and system output. These were described in either the paper, on the author's GitHub page, or in code documentation. We do observe differences in relation to 'system variation' which is inherent to training neural networks, where identical results are seldom obtained. These variations were small for most experiments, except for the 5 random seeds that led to substantial under-performing on one topic for BERT-large.

When looking at the second dimension of reproducibility defined by Cohen et al. (2018) (**findings**), we observe that BERT-base and BERT-large indeed clearly outperform the LSTM baselines from Stab et al. (2018) as well as our own stronger SVM+tf-idf non-BERT model on the stance detection task. We were able to reproduce the reported increase in performance of BERT-large over BERT-base and non-BERT models. However, BERT-large also showed considerable under-performance on one topic in 5 out of 10 seeds. We see this outcome as a confirmation that it is important to look at different seeds, and that care should be taken when drawing conclusions based on minor differences when working with neural models.

The third dimension of reproducibility is that of **conclusions**. Reimers et al. (2019) conclude that BERT strongly outperforms previous results on identifying arguments for unseen topics, which we confirm, and that these results are "very encouraging and stress the feasibility of the task" (Reimers et al., 2019, p. 575). The remainder of this paper provides further analyses to investigate whether our results also lead to this overall conclusion. In particular, we investigate how our models perform on individual topics (Section 4) and generic topic-independent signals in the data (Section 4.3).

4 Topic Specifics in Classification

To support the conclusions in Reimers et al. (2019) on the success of cross-topic stance detection, we expect a relative stability of performance over topics. The following sections go into some details not explored in Reimers et al. (2019), specifically the cross-topic performance of different topics, and the interaction between topic and class and its influence on performance.

4.1 Variance over (classes in) topics

Table 3 presents the performance of the models on individual topics. The results show that some topics perform considerably worse than others with the cross-topic training method (training on seven topics and testing on the held-out eighth topic). The *cloning* topic performs more than .07 F1 higher than the averaged model performance (F1 = .693 vs F1 = .617). The *abortion* and *gun control* topics perform almost .09 lower than the averaged model performance (F1 = .533 & .530 vs F1 = .617). Note that a difference nearing .10 in F1 score is relatively large, as it is comparable to the difference between the SVM performance and the state-of-the-art BERT models in the previous section.

A per-topic analysis in Table 3 shows that the SVM+tf-idf model performs within .10 points of the BERT-base model for seven of the eight topics, with some performing less than .3 points lower than BERT. The only exception is the topic *marijuana legalization*, which performs .28 points lower than the BERT model. The large average performance increase (+.11 in F1) over SVM comes from BERT-base improving performance on this one topic.

Figure 1 presents the BERT-base in-class F1 score of the three classes ('pro', 'con', 'no argument'), and in-topic averaged F1. The red line indicates the average model performance of .617. We see some consistency, e.g. the 'no argument' class consistently scoring around F1 = .80, but we also see some topic-specific behavior. *Cloning*, *minimum wage*, and *school uniforms* obtain higher F1 performance than average for all classes. In contrast, *death penalty*, *gun control*, and *abortion* perform considerably lower than the average F1 performance in the 'pro' and 'con' classes. These topics see in-class performance of even F1 < .50.

Each cross-topic model is trained by removing one topic from the training data. In this way, we remove a different number of training examples each time. The topics with the most training examples for a class (e.g. 'pro' in the *gun control* topic) therefore have a smaller training set for this class when training a cross-topic model. If there were a linear relationship between dataset size and performance, one would expect that topics with *fewer* training examples (and therefore more training examples left when this topic is left out of training) to do better than topics with *more* training examples (whose cross-topic models lose more training examples).

Table 1 does show that the 'no argument' class has a three times larger proportion of the training set than the 'pro' and 'con' classes, which could explain the better performance of this class in all topics, but training set size difference does not account for the between-topic variation in the 'pro' and 'con' classes. Instead, Table 1 shows that topics with the most training examples (that means, the largest set of examples removed in a cross-topic model) do not have the worst performing cross-topic models in Figure 1. For example, the *abortion* topic has relatively few 'con' examples removed (591) compared to other classes such as *cloning*, *death penalty*, and *nuclear energy*, and yet has the lowest in-class F1 for the 'con' class (in-class F1 = .40). Performance thus appears to be less related to the number of training examples.

We investigated the source of low performance on the 'pro' and 'con' class in the *abortion* topic with confusion matrices, and compared this to a topic where pro and against did not under-perform (*minimum wage*). We did not pick one specific seed, but calculated the mean percentage of 'true' examples in each confusion matrix cell over all 10 seeds. In the *abortion* topic, 44 % of 'pro' arguments get classified as 'against', and only 33% get correctly classified as 'pro'. The *minimum wage* topic shows no discernible pro/against classification confusion, and 60% of all true 'pro' and 'against' arguments are correctly classified. The section below analyzes the misclassifications in low-performing topics.

4.2 Qualitative Analysis of Misclassification

The low performance of 'pro' and 'con' in some topics (*abortion*, *gun control*, and *death penalty*) warrants some further investigation. Table 5 shows four example misclassifications between 'pro' and 'con' by BERT-large in the test examples the model encountered on these topics.[9]

[9]To ensure we are not cherry-picking examples, we looked at errors that were not unique to just one seed, and identified these examples as salient examples of a general trend.

held-out topic	abortion	cloning	death penalty	gun control	marijuana legalization	minimum wage	nuclear energy	school uniform
SVM+tf-idf	.463	.585	.482	.515	.323	.615	.598	.576
BERT-base	.533 (.011)	.693 (.013)	.562 (.012)	.530 (.013)	.607 (.016)	.670 (.009)	.660 (.011)	.678 (.016)
diff.	+.070	+.108	+.080	*+.028*	**+.283**	+.055	+.0850	+.102

Table 3: BERT-base's performance in F1 (macro) on different held-out topics. The *italicized* difference shows the smallest difference between the SVM model and the BERT-base model (on the gun control topic), while the **bolded** difference shows the largest difference (on the marijuana topic).

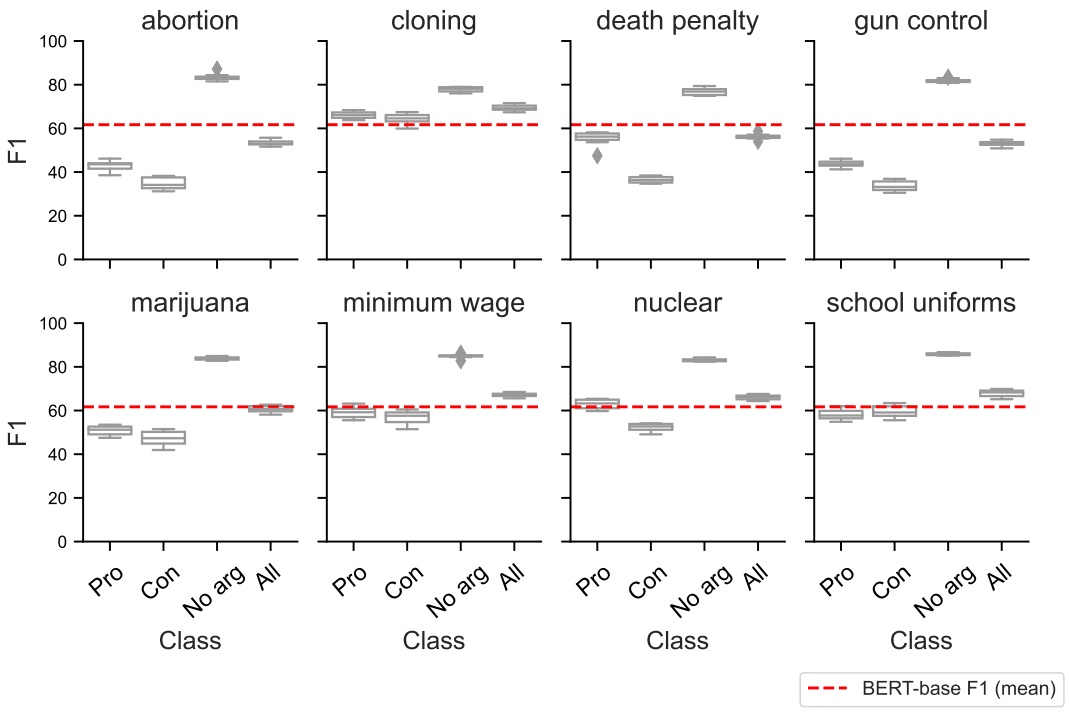

Figure 1: BERT-base's performance on different topics plotted in a boxplot, with on the y-axis the F1 score of the 4 categories plotted on the x-axis: 'pro', 'con', 'no argument', and overall. A longer boxplots means more variability over seeds in score. The red line represent the averaged F1 score of the same model (BERT-base), presented as model performance in Reimers et al. (2019).

We find two types of misclassifications, each related to topic-specific differences to stance classes. The first type is **misclassification due to the socio-cultural background knowledge and context** of a specific topic's arguments. The second type is related to a model taking the **stance towards a subcomponent of a topic** and confusing it for the text's overall stance on the topic, e.g. statements in the 'pro' class mostly expressing views *against* something else related to the argument (unwanted pregnancies, gun violence, innocents dying).

Examples of both issues are arguments centering around "many innocents (babies, children, mentally ill) will die". There are 5 variations of this argument in these 3 topics: row 1 and row 3 (*gun control*), row 8 (*abortion*), and rows 9 and 12 (*death penalty*) in Table 5. Not only is one usage of this argument traditionally connected to the 'pro' class of one topic (*gun control*), and the 'con' class of another (*abortion*), the implication is: innocents dying is bad. The model seems to lack this world knowledge, and for instance classifies this argument as 'pro' *death penalty*.

Another salient example is row 2 of Table 5. This argument argues *in favor* of gun rights for self-defense, but the model misclassifies this as *against* gun control. The model also fails to connect the second amendment discussion to the *against gun control* class. This is the same mistake made by the LSTM-model in Stab et al. (2018, p.3671), showing that BERT appears to not improve over LSTM on the topic-specific nuances here. In other words, it fails to correctly identify the socio-cultural dimensions (Du Bois, 2007) of this debate.

all topics				abortion topic			
Pro (vs Con)	Con (vs Pro)	No Argument		Pro (vs Con)	Con (vs Pro)	No Argument	
		vs Pro	vs Con			vs Pro	vs Con
pejorative	*morality*	basic	pronounced	seek	babies	way	anti
pronounced	format	section	threatens	illegal	abortion	against	ways
activity	*bill*	take	additional	reproductive	life	we	over
relations	workshop	**robert**	revolt	simply	conception	side	always
additional	workers	introduced	now	humane	simply	justify	thing
unexceptional	*sources*	unquestioned	proper	bear	risks	experience	question
threatens	*philosophical*	revolt	typical	lifers	abortions	held	performed
variable	coincidentally	scientifically	mentor	mother	complications	tell	debate
39th	*statutes*	**lifenews**	sharing	healthy	birth	single	illegal
where	phrases	individuals	denuded	lives	kill	had	equal

Table 4: Top-features for different topics according to SVM, Pairwise F-based feature analysis. We see potentially meaningful words in *italics* (the 'con' class has features based on morality and legality, e.g. bills and statutes), and potential spurious features in **bold** (such as names websites and even of individuals).

Topic	**True**	**Pred**	**Sentence**	**Frequency in seeds**
gun contr.	pro	con	"When high-capacity magazines were used in mass shootings, the death rate rose 63 % and the injury rate rose 156 % ."	9/10
gun contr.	con	pro	"[..] The Second Amendment protects an individual right to possess a firearm unconnected with service in a militia , and to use that arm for traditionally lawful purposes , such as self-defense within the home . "	7/10
gun contr.	pro	con	"In this crossfire , bullets would likely hit civilians (imagine a room filled with a crowd and three people shooting at each other) and the casualty count would increase."	9/10
gun contr.	con	pro	"Gun enthusiasts understand the benefit of large ammo feeders and wish to defend them because they recognize the advantage that such feeders give."	7/10
abortion	pro	con	"Not only has the biological development not yet occurred to support pain experience , but the environment after birth , so necessary to the development of pain experience , is also yet to occur ."	4/10
abortion	pro	con	"Warren concludes that as the fetus satisfies only one criterion, consciousness (and this only after it becomes susceptible to pain) , the fetus is not a person and abortion is therefore morally permissible ."	5/10
abortion	con	pro	It is argued that just as it would not be permissible to refuse temporary accommodation for the guest to protect him from physical harm , it would not be permissible to refuse temporary accommodation of a fetus .	2/10
abortion	con	pro	"92 % of abortions in America are purely elective – done on healthy women to end the lives of healthy children."	3/10
death pen.	con	pro	Mentally ill patients may be put to death .	2/10
death pen.	con	pro	Evidence shows execution does not act as a deterrent to capital punishment.	9/10
death pen.	pro	con	A system in place for the purpose of granting justice can not do so for the surviving victims , unless the murderer himself is put to death .	8/10
death pen.	con	pro	CON : " ... Since the reinstatement of the modern death pen. , 87 people have been freed from death row because they were later proven innocent .	9 /10

Table 5: Misclassifications on political topics with considerable 'pro' and 'con' confusion

4.3 SVM and Lexical Features

To analyze which words are used in relation to specific stances and topics, we trained an SVM model with tf-idf features on stance detection on all topics (F1 = .573). For each class pair ('pro' vs 'con', 'pro' vs 'no-argument', etc.), we extracted top-10 features with the highest coefficient for that specific class.

Table 4 presents the most important features of the topic-agnostic model trained on all topics. Some unigrams appear meaningful for the class. For instance, in the cross-topic setting, the word "morality" is a feature for the 'con' class. In contrast, the 'no argument' class is often identified with words that appear to have little content-relationship to the class identity: a topic-specific pro-life website (lifenews) or someone's name ('robert').

We also trained within-topic models to find whether there is topic-specific vocabulary related to stance that differs from the topic-agnostic model. Table 4 also presents the 10 most informative features for a model trained on only the abortion topic (F1 = .595). Immediately we see that there is only limited overlap with the lexical features used to decide between 'pro' and 'con' in a multi-topic scenario. Within only the abortion topic, the 'pro' and 'con' class are defined by concepts related to the lexical content of this specific discussion: babies, life, and birth. We also see the contrast between 'pro' arguments talking about reproduction and the mother, while the 'con' arguments mention life, conception, and babies. This lexical feature analysis shows no apparent overlap between the topic-specific features in the abortion model and the topic-independent features in the topic-agnostic model. This might indicate that vocabulary is quite specifically related to topics in stance detection.

5 Conclusion: Topic Matters

Stance detection is a difficult NLP task. Despite recent advances by pre-trained Transformers, these models have similar issues in a cross-topic setting as earlier models. This paper reproduced stance detection experiments with pre-trained Transformers by Reimers et al. (2019), training on seven topics and testing on an eighth topic. We found similar results, but also both class and topic influencing performance. Cross-topic BERT models perform below mean model performance in some topics (*abortion*, *gun control*) on the pro and con classes.

This makes us pause about Reimers et al. (2019)'s main claim: does BERT improve cross-topic stance detection over non-Transformer models? We argue this claim needs an asterisk: this cross-topic approach does not work as well for all topics. Different topics show specific vocabularies and socio-cultural contexts, and especially these specific contexts BERT cannot navigate. BERT models still make similar mistakes on gun control as the LSTM-based models in Stab et al. (2018).

These findings lead us to two take-aways. Firstly, we hypothesize that models like BERT rely more on topic-specific features for stance detection than topic-independent lexical words related to argumentation. Thorn Jakobsen et al. (2021) also recently found this, and connected BERT's cross-topic stance detection performance to its focus on spurious topic-specific lexical features ("gun", "criminal") rather than words related to argumentation. They also conclude a fair real-world evaluation of cross-topic stance detection means reporting the worst performing cross-topic pair rather than average performance over topics.

Secondly, we also think it is necessary to analyze the context of topics, and its relation to other debate topics within and outside the dataset. Most topics in stance detection studies are currently U.S. socio-political issues. This goes beyond a limitation of language, such as a focus on English without specifying this (Bender, 2019), since the same socio-cultural topics are not even universally relevant in the English-speaking world (gun control is not a salient discussion in Scotland). Such a focus on topic diversity is also important for use-cases. For diversity of viewpoints in search (Draws et al., 2021) or news recommendation (Reuver et al., 2021), stance detection needs to work on many different topics.

Schlangen (2021) states that we need to carefully define specific NLP tasks and capabilities needed to solve them. Modelling cross-topic stance detection in a topic-agnostic manner, while divorcing it from socio-cultural context, might not do justice to stance detection. Future work might focus on the specifics of topics: analyzing similarity between discussions (Wei and Mao, 2019), or modelling required socio-cultural contextual knowledge ('second amendment is related to gun control'). Models able to deal with topic-specific vocabulary and socio-cultural context of debates might improve on the state-of-the-art of cross-topic stance detection.

Acknowledgments

This research is funded through Open Competition Digitalization Humanities and Social Science grant nr 406.D1.19.073 awarded by the Netherlands Organization of Scientific Research (NWO). Our computing was done through SURF Research Cloud, a national supercomputer infrastructure in the Netherlands also funded by the NWO. We would like to thank dr. Nils Reimers for sending us their paper's data. We would also like to thank the anonymous reviewers, whose very helpful comments improved the paper. All opinions and remaining errors are our own.

References

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4756–4767, Online. Association for Computational Linguistics.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Mihaela Bošnjak and Mladen Karan. 2019. Data set for stance and sentiment analysis from user comments on croatian news. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 50–55.

K Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany J Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2018, page 156. NIH Public Access.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tim Draws, Nava Tintarev, and Ujwal Gadiraju. 2021. Assessing viewpoint diversity in search results using ranking fairness metrics. *ACM SIGKDD Explorations Newsletter*, 23(1):50–58.

John W Du Bois. 2007. The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 164(3):139–182.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

Antske Fokkens, Marieke Van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701.

Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. A survey on stance detection for mis-and disinformation identification. *arXiv preprint arXiv:2103.00242*.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2010. Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 conference short papers*, pages 253–257.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Stefan Ollinger, Lorik Dumani, Premtim Sahitaj, Ralph Bergmann, and Ralf Schenkel. 2020. Same Side Stance Classification Task: Facilitating Argument Stance Classification by Fine-tuning a BERT Model. *arXiv:2004.11163 [cs]*. ArXiv: 2004.11163.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Myrthe Reuver, Antske Fokkens, and Suzan Verberne. 2021. No NLP task should be an island: Multidisciplinarity for diversity in news recommender systems. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 45–55, Online. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance Detection Benchmark: How Robust is Your Stance Detection? *KI - Künstliche Intelligenz*.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec, Singapore. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Benno Stein, Yamen Ajjour, Roxanne El Baff, Khalid Al-Khatib, Philipp Cimiano, AG Semantic Computing, and Henning Wachsmuth. 2021. Same side stance classification. Preprint.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, page 327–335, USA. Association for Computational Linguistics.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1173–1176.

Lixing Zhu, Yulan He, and Deyu Zhou. 2019. Hierarchical viewpoint discovery from tweets using bayesian modelling. *Expert Systems with Applications*, 116:430–438.